Alignment of Molecules by the Monte Carlo Optimization of Molecular Similarity Indices

MARTIN F. PARRETTI, ROMANO T. KROEMER, JEFFREY H. ROTHMAN, W. GRAHAM RICHARDS

Physical and Theoretical Chemistry Laboratory, Oxford University, South Parks Road, Oxford OX1 3QZ, United Kingdom

Received 23 April 1996; accepted 22 February 1997

ABSTRACT: 3D-QSAR uses statistical techniques to correlate calculated structural properties with target properties like biological activity. The comparison of calculated structural properties is dependent upon the relative orientations of molecules in a given data set. Typically molecules are aligned by performing an overlap of common structural units. This ''alignment rule'' is adequate for a data set, that is closely related structurally, but is far more difficult to apply to either a diverse data set or on the basis of some structural property other than shape, even for sterically similar molecules. In this work we describe a new algorithm for molecular alignment based upon optimization of molecular similarity indices. We show that this Monte Carlo based algorithm is more effective and robust than other optimizers applied previously to the similarity based alignment problem. We show that QSARs derived using the alignments generated by our algorithm are superior to QSARs derived using the more common alignment of fitting of common structural units. *J Comput Chem* **18:** 1344–1353, 1997

Keywords: molecular alignment; molecular similarity; Monte Carlo; QSAR; CBG steroids

Introduction

Most 3-dimensional quantitative structure activity studies require that the molecules under investigation are aligned in space. This can be done by making a least squared fitting of atoms judged to be equivalent in separate structures. More generally one can adjust the relative positions of two molecules being aligned to optimize the similarity between them. This similarity may be calculated in terms of almost any property, but molecular shape and electrostatic potential are the most favored properties. Their preeminence arises from the fact that they are directly related to the noncovalent interactions (steric and electrostatic

Correspondence to: M. Parretti

 CCC 0192-8651 / 97 / 111344-10

interactions) thought to be chiefly responsible for biological effects at the molecular level.[1]

Molecular similarity indices provide a quantitative measure of the similarity of two compounds.[2,3] These indices are sensitive to the relative alignment of molecules.[4] Similarity optimization is performed in the belief that any two molecules that bind to a given receptor will do so with the relative alignment to which the molecules are most similar.

Generating an alignment of molecules by optimizing their similarity has several implications. First, the optimized similarity indices can be directly correlated with activity.[5–9] In this approach, each compound is compared for similarity with every other in the data set. Second, optimized similarity indices can be used as a basis for searching molecular data bases.[10] The search can be performed by screening out structures with a similarity index that falls below a given threshold. Third, the alignment generated by optimizing the similarity can be the basis for 3D-QSAR methods, such as comparative molecular field analysis (COMFA),[1] or it can be used for exploring a putative receptor or generating a pharmacophore model.

The most commonly used similarity indices are those of Carbo et al.[2] (R_{AB}) and Hodgkin et al.[3] (H_{AB}). The Carbo similarity of two molecules A and B, with respect to a given molecular property P, is assessed by

$$R_{AB} = \frac{\int P_A P_B d\nu}{(\int P_A^2 d\nu)^{1/2} (\int P_B^2 d\nu)^{1/2}},$$

where P_A and P_B are the properties of molecule A and molecule B, respectively. These two indices do not form an exhaustive list. A comprehensive review of molecular similarity and its diverse applications in chemistry was written by Rouvray.[11] This review covers not only historical aspects of the subject but also considers future directions. A lot of work was also performed recently in the development of quantum mechanical similarity indices (QMSI)[12–14] and novel and diverse definitions of similarity based upon these QMSIs.

The introduction of analytical Gaussian functions has proved to be invaluable for the fast and accurate computation of similarity integrals. For molecular electrostatic potential (MEP), Gaussian functions are fitted to the 1/r term of the MEP,[15]

$$P_r = \sum_{i=1}^{n} \frac{q_i}{(r - R_i)},$$

where the MEP P_r is calculated at a point r in space and q_i is the charge on atom i with nuclear coordinates R_i. As far as shape is concerned, Gaussians can be fitted to the electron density functions determined by taking the square of atomic wave functions, for example, STO-3G[16] functions.[17] The speed increase due to the implementation of the Gaussian approximation makes iterative calculations based on similarity a reasonable prospect.

A number of methods already exist for alignment. In the context of QSAR studies alignments are often performed by making a least squares fitting (LSF)[1,18] of atoms in common molecular structural units. Other possible alignments include combinations of center of mass and dipole and quadrupole overlap. When Carbo or Hodgkin indices are required, similarity optimization is often performed by the simplex method[19] and less commonly by gradient descent techniques.[20]

These methods have a number of weaknesses. While LSF of common structural atoms proves to be a good alignment technique for sets of structurally similar molecules, it is difficult to apply to groups of structurally more diverse compounds. In addition, alignments based on ESP are rarely as intuitive as those based on shape. Similarity based alignments lend themselves to the case of structurally varied molecules. However, the simplex and gradient descent methods also have failings that limit the usefulness of similarity based optimal alignments. Most optimization techniques tend to converge on local rather than global optima. The simplex and gradient methods are particularly prone to premature convergence.

On this basis we propose an alignment technique based upon the Monte Carlo optimization of molecular similarity indices. We will show that this optimizer displays superior performance to the simplex and gradient methods.

The Monte Carlo optimizer was tested on two data sets. First methotrexate (MTX) was aligned with dihydrofolate (DHF). The alignment of these two molecules is a classic problem.[21] Indeed these two molecules and other DHF reductase inhibitors have formed the basis of a very large number of modeling studies from QSAR to molecular docking. The second data set comprises 31 steroids known to bind to corticosteroid binding globulin (CBG). This data set has been used to test a range of QSAR techniques and as such provides a useful benchmark.[1,7,22] The data sets used by Cramer et al.[1] and Good et al.[7] contain errors and a corrected version was used in this study.[22] The data sets are

given in Figure 1a, b. Two different types of tests were applied to the second data set. First a direct comparison of the optimal similarity results obtained by simplex and Monte Carlo optimizers was made. Then the optimal alignments obtained by the Monte Carlo algorithm were used as the alignment rule in a quantitative structure activity relationship. The results of the QSAR were compared with those derived using other alignment rules.

QSARs were determined for the second data set using the similarity matrix methodology. This involved a partial least squares (PLS) analysis[23] of an $N \times N$ matrix of similarity values, and the corresponding SAR was derived. The results of the analyses were compared with those using other techniques from the literature.

Computational Methods

A Metropolis Monte Carlo based algorithm was implemented as our alignment optimizer. The Metropolis method can be broadly described as having the following features: the modification of the conformation of the system by generating random moves; the acceptance of every random conformational shift that results in an improvement of the property being optimized; and the acceptance of a small proportion of moves that do not correspond to an improvement of the property, subject to some acceptance criterion. This final feature of the Metropolis method is essential because it allows it to escape local minima and therefore to search a greater volume of similarity space.

Moves that correspond to a less favorable similarity index are accepted only if the difference in the latest and previously accepted similarity indices does not exceed a threshold value. Then an exponential of this difference is tested by comparison with a random number chosen between 0 and 1. The exponential term used is

$$\exp(-Cx\Delta),$$

where C is a constant that determines the severity of the acceptance criterion and D is the difference in the similarity numerators. The move is accepted if the exponential term is greater than the random number.

The size of permissible random translational jumps is limited to a maximum value. However, its choice has a significant influence on the performance of the algorithm. With this fact in mind, the maximum allowed jump is modified during the course of the optimization such that the total number of accepted moves is kept at approximately half the total number of moves performed.

For the purposes of this implementation, convergence was defined as a maximum number of moves without further improvement over the best value of similarity obtained.

With the realization that the Monte Carlo optimizer would not be a truly global optimizer, a

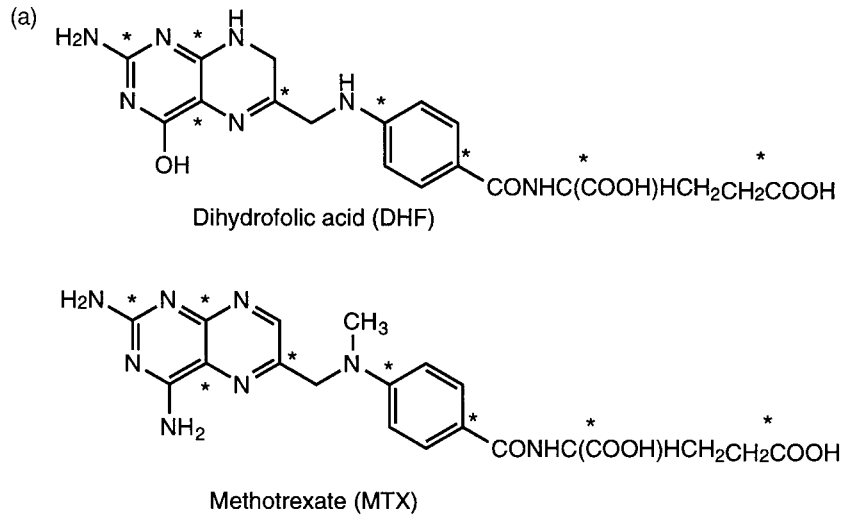

FIGURE 1. (a) Structures of dihydrofolate and methotrexate. (b) Structures of the steroid data set used in the analyses. (*) Atoms used in RMS fitting of the structures.

(b)

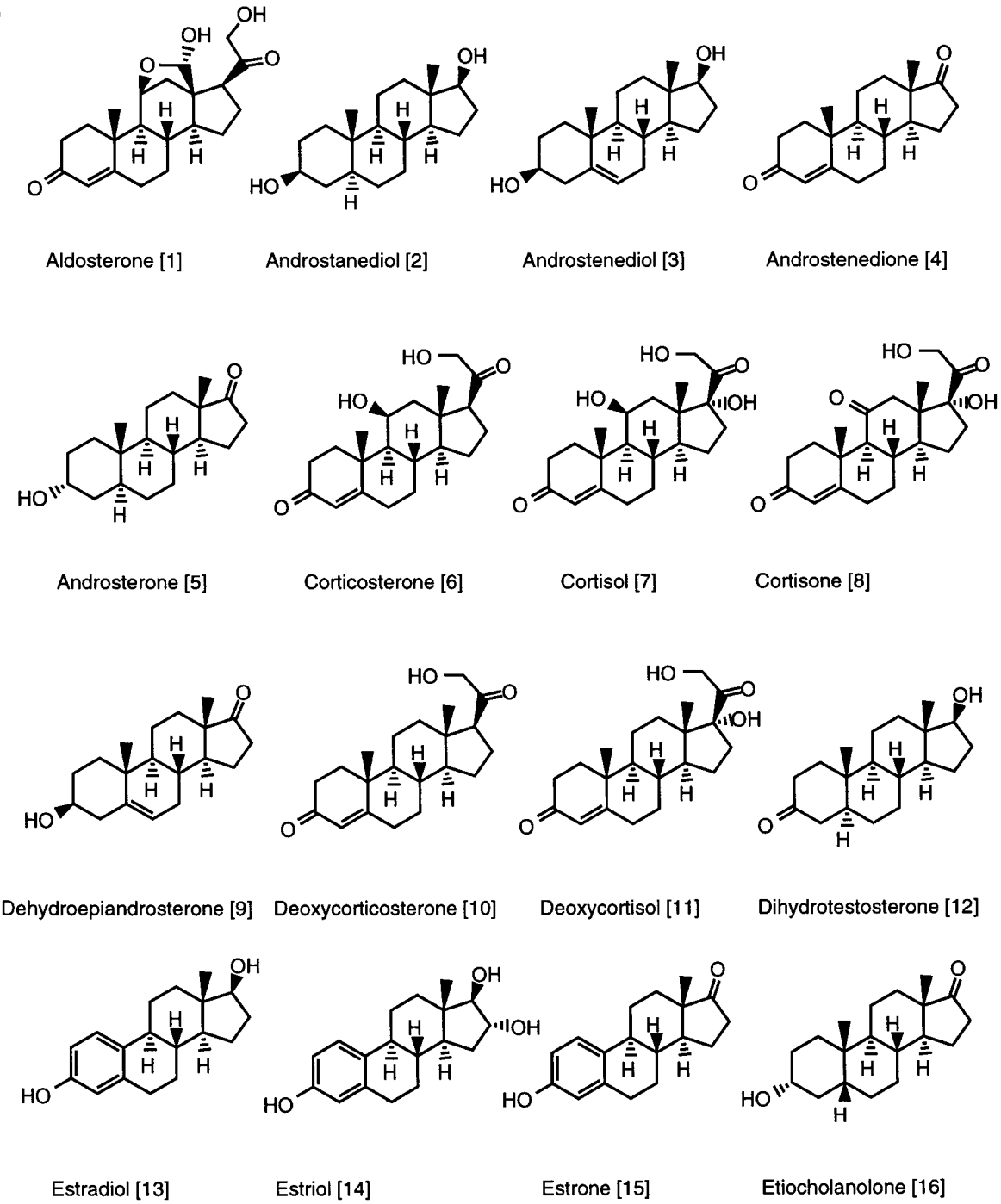

FIGURE 1. (Continued).

multiple trajectory option was included in the software. This feature allows the algorithm to restart from a random spatial orientation upon convergence of the previous trajectory.

The MTX and DHF models were constructed starting with coordinates taken from the Cambridge Crystallographic Data Base.[24] These structures and the corrected steroids were minimized using the standard Tripos[25] force field. Point charges for the corrected structures were calculated by RATTLER software.[26,27] These charges were fitted to the ESP calculated semiempirically

(b)

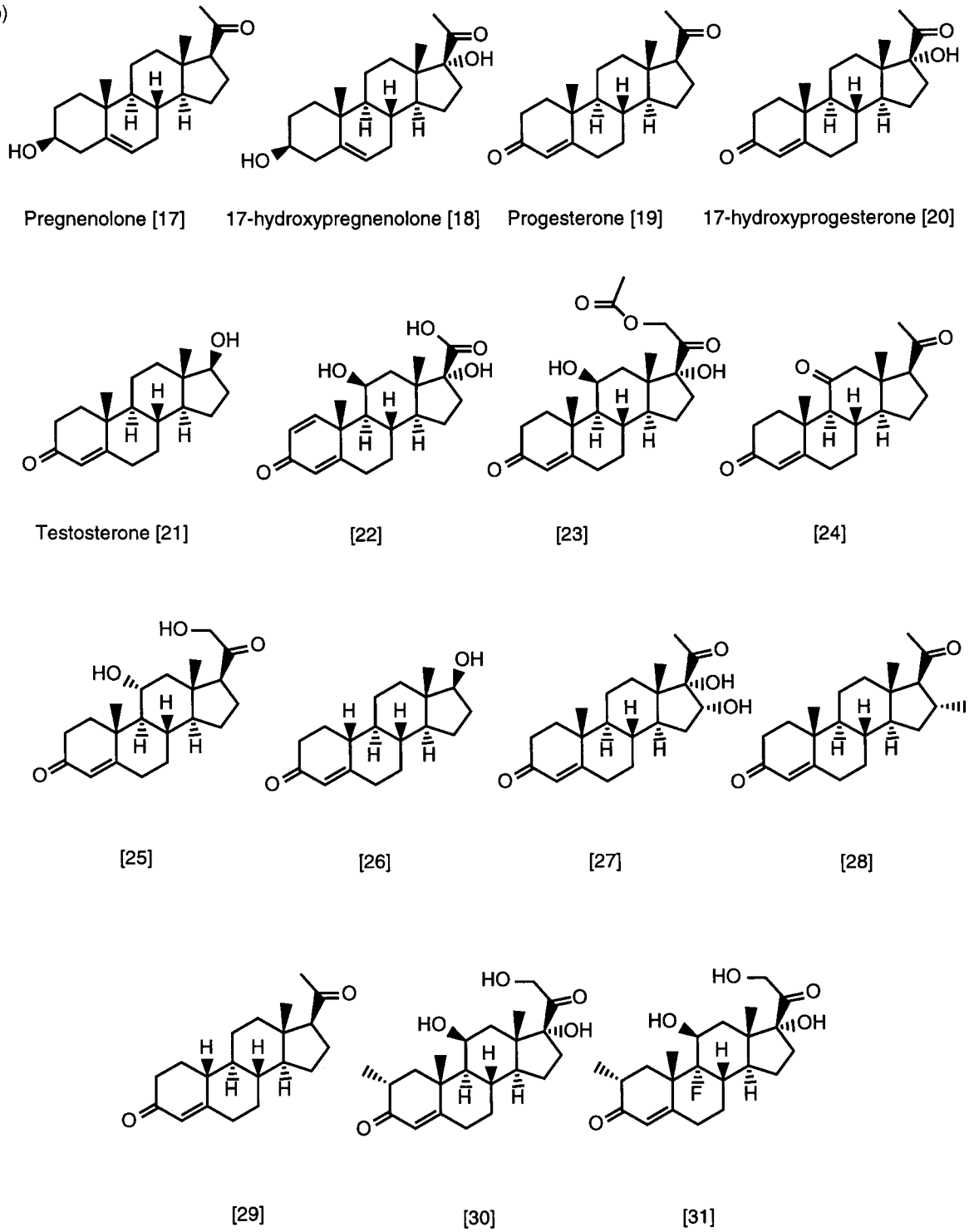

FIGURE 1. (Continued).

using the AM1[28] Hamiltonian implemented in MOPAC 6.0.[29] The choice of method of point charge calculation was recently found to have a significant effect on QSAR results.[30,31] ESP fitted charges at the semiempirical level of theory were found to perform very well in QSAR and hence are used here.

We used the PLS algorithm implemented in TSAR.[26] All cross-validated runs used a leave one out cross-validation strategy.

Results and Discussion

Two initial alignments of MTX and DHF molecules were used to test the optimizer. First MTX was rotated by 90° about the axis perpendicular to the benzene moiety. In the second case MTX was rotated by 180° about the same axis and 180° about the axis described by the two substituent bonds of the benzene ring. Electrostatic alignments were then performed. The optimized similarity indices were 0.68 and 0.66, respectively. The molecules were then aligned by a root mean squares (RMS) fitting of the atoms marked in Figure 1a. The electrostatic similarity calculated at this alignment was 0.64, which was actually a little lower than for the ESP optimizations. The alignment of MTX at its optimum ESP similarity was very slightly skewed from the minimal separation of nuclear centers of the RMS alignment.

The results of all ESP optimizations that were performed on the steroids data set are given in Table I. This table lists the results of the optimization of the similarity index between steroid 11 and all the other steroids in the data set. Alignment to steroid 11 was made because this steroid has the

TABLE I.
MEP Carbo Similarity Indices for Steroid Data Set Using Monte Carlo and Simplex Optimization Methods.

	MEP Carbo Similarity Index			
Steroid	Monte Carlo Common Atom Fitted	Simplex Common Atom Fitted	Monte Carlo Random orientation	Simplex Random Orientation
1	0.727	0.725	0.723	0.725
2	0.552[a]	0.539	0.582	0.612
3	0.575	0.581	0.576	0.589
4	0.679	0.658	0.685	0.640
5	0.643	0.329	0.645	0.647
6	0.817	0.816	0.793[b]	0.420
7	0.936	0.938	0.915[b]	0.544
8	0.842	0.846	0.817[b]	0.846
9	0.577[b]	0.472	0.638	0.387
10	0.904	0.905	0.902	0.485
11	1.000	1.000	1.000	1.000
12	0.672	0.641	0.669	0.463
13	0.579	0.551	0.581	0.573
14	0.581	0.542	0.620	0.617
15	0.599	0.446	0.592	0.539
16	0.644	0.348	0.649	0.669
17	0.673	0.635	0.667	0.626
18	0.772	0.750	0.769	0.473
19	0.763	0.759	0.765	0.680
20	0.851[b]	0.850	0.849	0.570
21	0.650	0.558	0.644	0.606
22	0.897	0.903	0.908[b]	0.903
23	0.807	0.801	0.799[b]	0.801
24	0.647	0.636	0.627[b]	0.636
25	0.845	0.549	0.847	0.549
26	0.650	0.580	0.649	0.580
27	0.766	0.648	0.763	0.649
28	0.677	0.679	0.585[b]	0.678
29	0.751	0.417	0.760	0.417
30	0.936	0.569	0.937[b]	0.567
31	0.917	0.574	0.924	0.574

Similarity optimization is with respect to steroid 11.
[a] Three trajectories required to get this result.
[b] Two trajectories required to get this result.

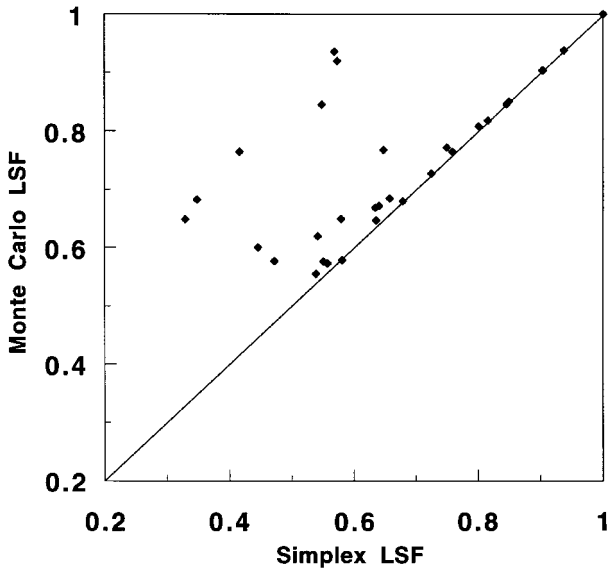

FIGURE 2. Plot of the optimal similarity indices for the steroid data set as calculated by the Monte Carlo versus simplex techniques, commencing from common backbone aligned initial orientations (LSF).

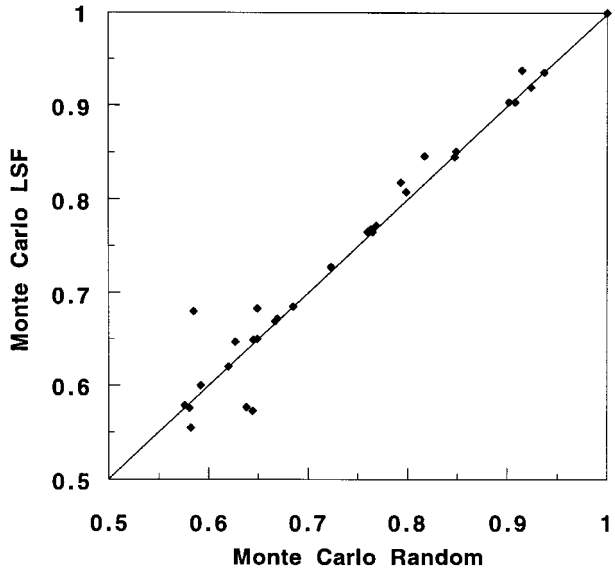

FIGURE 4. Plot of the optimal similarity indices for the steroid data set as calculated by the Monte Carlo algorithm: common backbone aligned initial orientation (LSF) versus randomized initial orientation.

highest biological activity in the data set. Columns 2 and 3 in Table I list the optimized similarity indices for optimization by the Monte Carlo and simplex methods, respectively, commencing from the common backbone fitted orientation. Surprisingly, comparison of these indices shows significant variation, as is graphically illustrated in Figure 2. The simplex method is expected to perform at its best with this prealignment, and yet the Monte Carlo technique still significantly outperforms it. Optimization of alignment was also performed commencing from randomized initial ori-

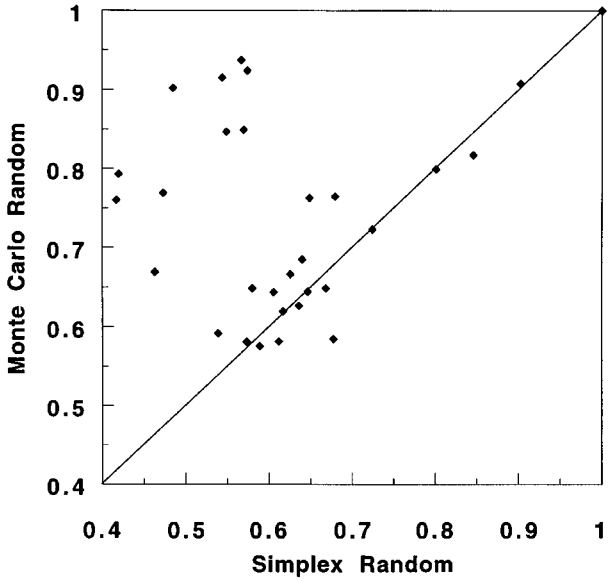

FIGURE 3. Plot of the optimal similarity indices for the steroid data set as calculated by the Monte Carlo versus simplex techniques, commencing from randomized initial orientations.

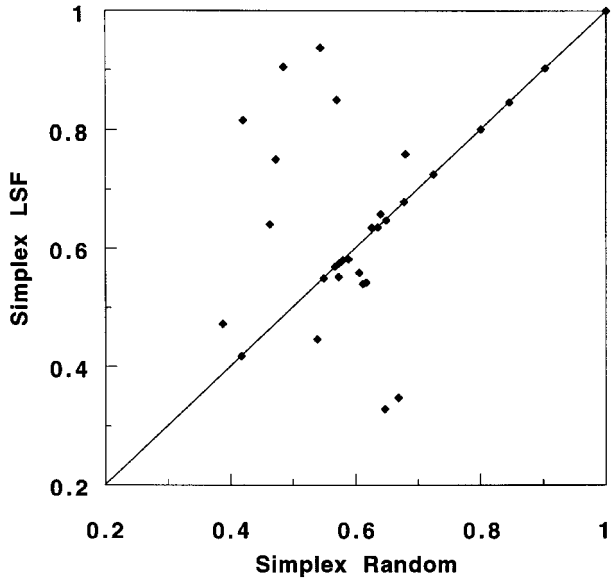

FIGURE 5. Plot of the optimal similarity indices for the steroid data set as calculated by the simplex algorithm: common backbone aligned initial orientation (LSF) versus randomized initial orientation.

entations. The resulting optimized indices are given in columns 4 (Monte Carlo) and 5 (simplex) of Table I. The graphical comparison of these optimizations is given in Figure 3. The aim of this calculation was to provide some indication of the relative robustness of the Monte Carlo and simplex routines. The Monte Carlo optimization again outperformed the simplex in a more substantial manner than from the common backbone aligned starting orientation. Figure 4 shows the high degree of consistency between the results of Monte Carlo runs that commence from the different initial orientations. There is very little consistency between the two simplex runs (Fig. 5).

The time penalty for Monte Carlo optimization is not very large, especially when the quality of the results is considered. On average a single trajectory Monte Carlo run took only about 40 s of CPU time on a Silicon Graphics Indigo (R4000) workstation. This compares to an average of about 10 s for the simplex algorithm. It is hoped that shape based similarity alignments can be carried out as well by the Monte Carlo algorithm.

The result of the QSAR analysis for the steroids is given in Table II. Results for previously determined QSARs on the same data set are also given, but some caution needs to be exercised in comparing the results.

The first set of QSAR results (labeled MC matrix) are the results of a QSAR determined for the Monte Carlo aligned steroids. The second set (LSF matrix) give the results for the common backbone aligned steroids as carried out by Good et al.[7] but with the corrected steroid structures. It is interesting that in the original work by Good et al.[7] the shape descriptor based QSAR was found to model the data better than the ESP based model, but the reverse was found to be true for the corrected data set. In both the Monte Carlo and the LSF aligned SARs the ESP based relationships modeled (conventional r^2) and predicted (cross-validated r^2) the data best. The Monte Carlo aligned SAR shows quite an improvement over the backbone fitted alignment model.

The third set of tabulated results is for QSAR models based on steroids 1–31 and 1–30, respectively. These results are included to provide a

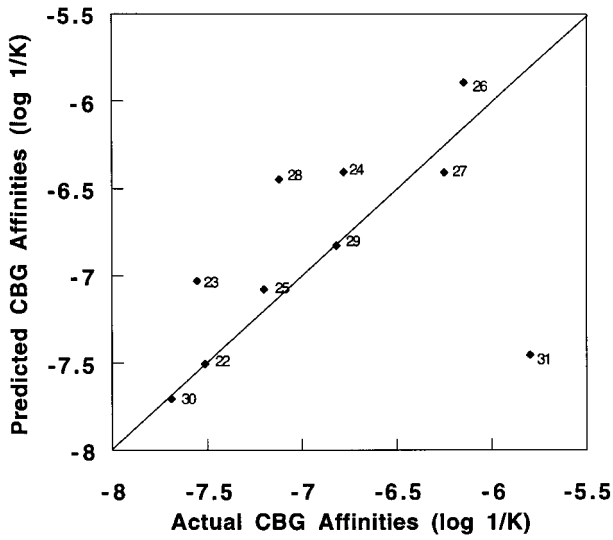

FIGURE 6. Plot of the QSAR predicted CBG binding affinity of steroids 22–31 versus their actual measured binding affinities.

TABLE II.
Results of QSAR Analyses.

Method	Cross-Validated r^2	Conventional r^2	No. Components
ESP MC matrix	0.780	0.819	1
Shape MC matrix[a]	0.451	0.569	2
ESP / shape MC matrix[a]	0.658	0.704	1
ESP LSF matrix	0.724	0.798	2
Shape LSF matrix	0.446	0.520	1
ESP / shape LSF matrix	0.724	0.800	1
ESP MC matrix 31	0.734	0.764	1
ESP MC matrix 30	0.820	0.841	1
Autocorrelation vector 31	0.630	0.820	—
Autocorrelation vector 30	0.840	0.920	—

[a]Here shape means shape similarity evaluated with the alignment derived from the molecular electrostatic potential similarity optimization.

comparison with the new autocorrelation vector QSAR results of Wagener et al.,[22] who chose an extended data set for modeling rather than the more typical use of steroids 1–21 for SAR derivation followed by prediction testing of the remaining steroids. Because the autocorrelation data set was unavailable at the time of performing the calculations, the data sets used in both QSARs were not identical, However, they are both structurally and stereochemically correct.

The choice of a 30 steroid SAR by Wagener et al.[22] was prompted by the fact that steroid 31, an outlier in the data set, appeared to be biasing the QSAR unfavorably. It is interesting to note that the 31 compound QSAR based upon our Monte Carlo alignments is apparently much less sensitive to this outlier, and consequently predictions of activity made using our SAR are closer to the observed values.

RMS data for our alignments have not been included in this article. We briefly explain why this decision was made. The alignment of MTX to DHF actually showed that the ESP alignments, both of which were consistent, resulted in a slightly higher similarity than the RMS alignment. This is a feature observed most strikingly for a number of the steroid structures, for which the optimal ESP alignments will often not coincide with a minimal overlap of the nuclear centers. Sometimes ESP and RMS alignments are not even qualitatively similar. One example is the ESP alignment of structures 5 and 11 (androsterone and deoxycortisol). The similarity at the optimal electrostatic alignment was 0.645. The corresponding similarity value calculated for minimal RMS alignment[1,7] was −0.092. These similarities are clearly very different, as are the corresponding alignments.

One of the major results of this study is that a QSAR derived from our electrostatic alignments models and predicts the activities of these molecules better than a QSAR for the same molecules aligned by RMS fitting. The RMS alignments are different from the optimal ESP alignments. Given these points it is difficult to imagine what help the RMS fit values can be.

Conclusions

We have outlined a method for aligning molecules with respect to their ESP. Previous techniques applied to this problem have proved to not be particularly robust. We have shown that our Monte Carlo based alignment tool is robust.

A demonstration of the utility of such an alignment tool has also been given. The molecular alignments that we have generated were used as the basis for a QSAR of the data set. Our results show an improvement over the ''classical'' alignment rule for this data set, which involves aligning common structural atoms. Comparison is also made with QSARs derived from autocorrelation vectors, which are orientation independent. Our results are very comparable for the extended data set model considered. The data sets differ in the methods used to calculate ESPs. The data set of Wagener et al.[22] uses the empirical PEOE method,[32] in contrast to the semiempirical method employed in this work. As a result, small differences in the results of these two different methods should not be studied too hard because of this difference in the data sets. It appears that our model is a little less sensitive to outlying structures, and as such predictions using our model are somewhat less easily biased.

Clearly this is not the end of the problem. The optimization technique needs to be extended to alignment based on steric descriptors, and also, ideally, on a combined steric/electrostatic descriptor, because steric and electrostatic factors together are thought to account for the majority of observed biological effects at the molecular level. Work is proceeding on these algorithms.

In summary, we have presented an algorithm that will successfully align molecules. We have demonstrated that such alignments may be usefully applied to deriving good QSARs and that the alignments are carried out at only a small time penalty as compared to the other algorithms mentioned.

References

1. R. D. Cramer III, D. E. Patterson, and J. D. Bunce, *J. Am. Chem. Soc.*, **110**, 5959 (1988).
2. R. Carbo, L. Leyda, and M. Arnau, *Int. J. Quantum Chem.*, **17**, 1185 (1980).
3. R. Carbo and L. Domingo, *Int. J. Quantum Chem.*, **32**, 517 (1987).
4. C. Burt and W. G. Richards, *J. Comput.-Aided Mol. Design*, **4**, 231 (1990).
5. C. Burt and W. G. Richards, *J. Comput. Chem.*, **11**, 1139 (1990).
6. A. Seri-Levy, R. Salter, S. West, and W. G. Richards, *Eur. J. Med. Chem.*, **29**, 687 (1994).
7. A. C. Good, S. J. Peterson, and W. G. Richards, *J. Med. Chem.*, **36**, 2929 (1993).
8. A. C. Good, S. So, and W. G. Richards, *J. Med. Chem.* **36**, 433 (1993).

9. R. Benigni, M. Cotta-Ramusino, F. Giorgi, and G. Gallo, *J. Med. Chem.*, **38**, 629 (1995).
10. A. C. Good, E. E. Hodgkin, and W. G. Richards, *J. Comput.-Aided Mol. Design*, **6**, 513 (1992).
11. D. H. Rouvray, *Topics Curr. Chem.*, **173**, 1 (1995).
12. E. Besalu, R. Carbo, J. Mestres and M. Sola, *Topics Curr. Chem.*, **173**, 31 (1995).
13. R. Carbo, E. Besalu, L. Amat, and X. Fradera, *J. Math. Chem.*, **18**, 2 (1995).
14. R. Carbo, E. Besalu, L. Amat, and X. Fradera, *J. Math. Chem.*, **19**, 47 (1996).
15. A. C. Good, E. E. Hodgkin, and W. G. Richards, *J. Chem. Informatics Comput. Sci.*, **32**, 188 (1992).
16. M. J. Frisch, M. Gordon, H. B. Schlegel, K. Raghavachari, J. S. Binkley, C. Gonzalez, D. J. Defrees, D. J. Fox, R. A. Whiteside, R. Seeger, C. F. Melius, J. Baker, R. Martin, L. R. Kahn, J. J. P. Stewart, E. M. Fluder, S. Topiol, and J. A. Pople, Gaussian 88, Gaussian Inc., Pittsburgh, PA, 1988.
17. A. C. Good and W. G. Richards, *J. Chem. Informatics Comput. Sci.*, **33**, 112 (1993).
18. M. S. Allen, Y.-C. Tan, M. L. Trudell, K. Narayanan, L. R. Schindler, M. J. Martin, C. Schultz, T. J. Hagen, K. F. Koehler, P. W. Codding, P. Skolnick, and J. M. Cook, *J. Med. Chem.*, **33**, 2343 (1990).
19. J. A. Nelder and R. Mead, *Comput. J.*, **7**, 308 (1965).
20. A. J. McMahon and P. M. King, *J. Comput. Chem.*, **18**, 151 (1997).
21. J. M. Blaney, C. Hansch, C. Silipo, and A. Vittoria, *Chem. Rev.*, **84**, 333 (1984).
22. M. Wagener, J. Sadowski, and J. Gasteiger, *J. Am. Chem. Soc.*, **117**, 7769 (1995).
23. S. Wold, C. Albano, W. J. Dunn, U. Edlund, K. Esbenson, P. Geladi, S. Hellberg, W. Lindeberg, and M. Sjøstrøm, In *Chemometrics*: *Mathematics and Statistics in Chemistry*, B. Kowalski, Ed., Reidel, Dordrecht, The Netherlands, 1984, p. 17.
24. Cambridge Crystallographic Database, University Chemical Laboratory, Cambridge, U.K.
25. J. G. Vinter, A. Davies, and M. R. Saunder, *J. Comput.-Aided Mol. Design*, **1**, 31 (1987).
26. Oxford Molecular Ltd., Oxford, U.K.
27. G. Ferenzcy, C. A. Reynolds, and W. G. Richards, *J. Comput. Chem.*, **11**, 159 (1990).
28. M. J. S. Dewar, E. G. Zoebisch, E. F. Healey, and J. J. P. Stewart, *J. Am. Chem. Soc.*, **107**, 3902 (1989).
29. J. J. P. Stewart, *MOPAC 6.0*, Quantum Chemical Program Exchange 455, 1990.
30. R. T. Kroemer and P. Hecht, *J. Comput. Chem.*, to appear.
31. M. S. Allen, A. J. LaLoggia, L. J. Dorn, M. J. Martin, G. Costatino, T. J. Hagen, K. F. Koehler, P. Skolnick, and J. M. Cook, *J. Med. Chem.*, **35**, 4001 (1992).
32. J. Gasteiger and M. Marsili, *Tetrahedron*, **36**, 3219 (1980).